ІНТЕГРАЦІЯ ФІЛОЛОГІЇ І ТЕХНІЧНИХ НАУК

УДК 658.012

***Е.А. ОРОБИНСКАЯ,*** аспирант, НТУ «ХПИ», Харьков;
Университет им. Братьев Люмьер Лион-2, Лион, Франция
***А.Ю. ДОРОШЕНКО,*** студент, НТУ «ХПИ», Харьков

# ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

У статті пропонується розглянути, що таке онтологія, яку вона має структуру, які існують підходи для роботи з онтологіями при автоматичній обробці текстів на природній мові, а саме створення, настройка та використання онтологій. Представлені проекти для роботи з онтологіями, зроблено висновки по даній темі.

В статье предлагается рассмотреть, что такое онтология, какую она имеет структуру, которые существуют подходы для работы с онтологией при автоматической обработке текстов на естественном языке, а именно создание, настройка и использование онтологии. Представлены проекты для работы с онтологией, сделаны выводы по данной теме.

In this paper we suggested to consider what is ontology, what structure it has, the actual approaches of ontology engineering and ontology usage towards the tasks of natural language texts treating are considered and analyzed. There are projects for work with ontology and conclusions.

**Введение**. Термин онтология пришел из философии, где он используется для обозначения системы знаний, которые относятся к окружающему миру, а также для работы с этой системой знаний. «Онтология – это явная спецификация концептуализации» [19]. Здесь концептуализация означает абстрактное представление предметной области. Распространено также определение: «Онтология – общее понимание некоторой области интереса» [17]. На сегодняшний день под онтологией можно понимать: надежный семантический базис в определении содержания; общую логическую теорию, которая состоит из словаря и набора утверждений на неко-

тором языке логики; основу для коммуникации между людьми и компьютерными агентами.

Таким образом, в настоящее время не существует общепринятого определения. Практически все модели онтологий, в той или иной степени, содержат концепты (понятия, классы, сущности, категории), свойства концептов (слоты, атрибуты, роли), отношения между концептами (связи, зависимости, функции) и дополнительные ограничения, которые помогают автоматически обрабатывать тексты [18].

Для представления элементов в предметной области определенного концепта, используется термин экземпляр. Онтология вместе с множеством отдельных экземпляров составляют базу знаний. Заметим, что, между онтологиями и базами знаний нечеткая грань.

Считается, что онтология представляет собой базу знаний, описывающую факты, которые предполагаются всегда истинными в рамках определенного сообщества на основе общепринятого смысла используемого словаря. База знаний же может описывать факты и утверждения, истинность которых зависит от состояния переменных внешней среды. В данном вопросе пока нет полной ясности. Как правило, онтологии, предназначенные для автоматической обработки текстов, выделяются в отдельную категорию лексических онтологий [5], которую, в свою очередь, можно разделить на две группы: т.н. общие (***top-level ontology***) и прикладные онтологии (***domain ontology***), т.е. относящиеся к определенной предметной области [1]. В данной работе будут рассмотрены существующие подходы к созданию и использованию онтологий обеих групп.

**1. Использование генеральной онтологии.** В информатике онтология рассматривается, прежде всего, как модель представления знаний. Онтология позволяет организовать в иерархическую структуру понятия (концепты), встречающиеся в тексте, и получить знания более высокого уровня (синонимы, гиперонимы, и т.д.), пригодные для последующего автоматического и ручного анализа.

Когда речь идет об использовании общей онтологии для обработки текстов, большинство авторов предлагает использовать Word Net [3]. Word Net – это общий словарь терминов английского языка, который группирует слова в систему (группы синонимов, создающих контекст), и который организует синсеты между этими группами, т. е. отношения гиперонимии, гипонимии, меронимии и оломинии.

Аналогичный проект для русского языка, известный как RussNet, разрабатывается с 1999 года исследовательской группой под руководством И. В. Азаровой (Кафедра математической лингвистики Филологического факультета Санкт-Петербургского государственного университета) [5].

Однако, использование WordNet в качестве онтологии, для представления (описания) текстов не так уж очевидно. Первая операция, как прави-

ло, состоит в том, чтобы извлечь отношения гипер- и гипонимии, которые можно рассматривать как обобщение или уточнение (*is-a* и *part of*).

Выполнение первого этапа позволяет далее решить две возможные задачи:

1 уточнить меру подобия двух текстов, учитывая семантическое расстояние между словами;

2 улучшить описание текста с помощью гиперонимов, т.е. более общих понятий.

Первое решение позволяет определить семантическое подобие на основе базового графа. Известно множество методов измерения семантического подобия текстов: метод Ph. Resnik, Wu&Palmer, Jiang&Conrath и др. [7].

Другое решение состоит в систематическом добавлении обобщающих понятий (гиперонимов) в описание текста, в котором находится интересующее слово [8]. Этот подход позволяет объединить тексты, которые не используют общий словарь, но на самом деле связаны единой тематикой. В данном случае основная проблема состоит в том, что каждое слово зачастую связано с несколькими разными смыслами (синсетами). Поэтому каждый раз необходимо принять решение: добавлять ли новые гиперонимы, соответствующие разным смыслам, или выбрать один из них, кажущийся наиболее вероятным. Это известная проблема снятия неоднозначности (***Word Sense Disambiguation* – WSD**), являющаяся далеко не тривиальной [9]. Данная проблема возникает в значительно меньшей степени, когда речь идет о специальных прикладных онтологиях, которые будут рассмотрены в следующей части работы.

С другой стороны, существует иная проблема определения необходимого уровня обобщения или детализации понятий, т.е. выбора количества гиперонимов/гипонимов. Достаточно ли, например, добавить родовое понятие «семейство собачьих», для слова «собака». Или стоит также добавить «хищник» (гипероним для «семейство собачьих»). На сегодняшний день не существует готового решения этой проблемы, которую периодически пытаются решить, развивая существующие методы.

**2. Разработка прикладных онтологий.** В литературе часто разделяют задачи создания онтологии и настройку (или развитие) существующей онтологии. Некоторые авторы предлагают использовать единое понятие управления онтологиями (***ontology engineering***), [2]. Авторы работы [12] сравнивают создание и настройку онтологии со слоеным пирогом (***Ontology Learning Layer Cake***), подразумевая, что весь процесс разбивается на отдельные этапы, каждый из которых можно рассматривать как задачу из области Text Mining.

***2.1. Построение прикладных онтологий.*** Существует несколько способов создания онтологий «от А до Я» на основе текстов, написанных на естественном языке. Один из первых методов был предложен Гельфандом и его коллегами (Gelfand et al, [10]). Идея метода заключается в том, чтобы

построить прикладную онтологию на основе уже существующей, но более общей онтологии. Авторы используют Word Net, для создания семантического графа (***Semantic Relationship Graph*** или ***SRG***). Одним из преимуществ такого подхода является то, что извлекаемые слова априорно имеют соответствующую прикладной области семантику, т.к. из Word Net выбирается подходящий синсет, что в значительной степени снимает проблему неоднозначности слов. Кроме того, они могут быть связаны с другими словами, которые не представлены в исследуемых текстах (т. наз. ***augmenting words***). Другим преимуществом использования такого подхода является то, что, как правило, часто существует несколько путей, связывающих два слова в графе, что гарантирует дополнительную устойчивость (защищенность) системы. Следующим этапом может быть кластеризация текста (или ***blocking***) для выделения фрагментов тесно связанных терминов.

Иной подход заключается в использовании иерархической структуры, которая строится либо по принципу ***top-down***, когда берется один общий концепт, объединяющий все понятия, которые затем постепенно уточняются; либо по принципу ***bottom-up***, когда сначала выбирается специальный концепт и затем он постепенно обобщается. В некоторых работах предпринимается собственный подход по обработке текстов на естественном языке (***Natural Language Processing, NLP***), где во внимание принимается синтаксис выражений [14]. Авторы этой работы разбивают текст на тематически однородные параграфы и затем работают с триплетами глагол-отношение-существительное, чтобы построить «структурированные области». Целью разработанной ими системы SVETLAN является обнаружение контекстно-зависимых слов в какой-либо предметной области.

И, наконец, последний подход к созданию онтологий – это использование матрицы концептов [15]. Основным недостатком такого подхода является то, что он достаточно дорогостоящ и плохо применим к большим областям. Документы представляются не при помощи классической сумки слов (***bag-of-words***), а их синсетами в WordNet и своими гиперонимами (авторы остановились на 4 уровнях). Следующий этап состоит в использовании технологии FCA – для сравнения онтологий, чтобы получить решение проблемы группировки концептов.

***2.2. Настройка онтологии.*** Настройка онтологии тесно связана с только что рассмотренной проблемой их построения. В рамках тематики SOAT (***Semi-automatic domain Ontology Acquisition Tool***) написано множество работ.

Некоторые авторы [11] предлагают объединить создание и настройку онтологий в общую проблематику – ***ontology engineering***. Они рассматривают эти этапы как циклический процесс постоянного совершенствования онтологии. Архитектура их системы, позволяющей реализовать задачу создания и настройки прикладной онтологии, состоит из таких модулей: модуль управления текстом, где осуществляется отбор ресурсов и методов

обработки, которые будут использованы впоследствии; модуль обработки текстов, основанный на системе SMES (***Saarbrücken Message Extraction System***), которая позволяет выполнять синтаксический анализ, а также выполняет другие операции (толкиенизацию, лексический анализ, распознавание собственных имен и т.д.); модуль настройки, который строит таксономию концептов, основанную на алгоритмах соответствия базовых понятий и регулярных выражений, выявляет общие закономерности связей между словами и т.д.; подмодуль OntoEdit – это подмодуль системы создания онтологий, который позволяет добавлять концепты, обнаруженные в существующей онтологии полуавтоматическим способом.

Авторы подчеркивают, что механизм настройки онтологий, используемый в их системе, не позволяет сразу автоматически создать совершенную онтологию, но помогает ее построить, выдавая рекомендации разработчику относительно того, как следует модифицировать настраиваемую онтологию.

В работах, выполненных на основе системы THESUS [13], также используется стратегия полуавтоматической обработки. Обрабатываемые документы были взяты из Интернета. В частности, были использованы гиперссылки для того, чтобы найти ключевые слова, наиболее релевантные описанию теста. Особенностью этих работ является использование общей онтологии (здесь Word Net) для сравнения web-документов и онтологии предметной области. Идея состоит в том, чтобы одновременно сопоставлять и ключевые слова web-документов, и концепты онтологии с графом Word Net. После этого, достаточно использовать модифицированные критерии схожести в самом графе. После того как тексты будут описаны, к ним можно будет применить алгоритмы кластеризации с тем, чтобы выделить однородные группы, после чего выполняется операция лейбеллинга (labelling). В последние годы этот алгоритм был усовершенствован [16]. Авторы основываются на этой технологии с тем, чтобы предложить возможные изменения в прикладной онтологии в зависимости от результатов кластеризации. Они распределяют слова текста между концептами, т.е. словами, которые находятся в генеральной онтологии (Word Net) и экземплярами (сущностями).

**Заключение.** В целом, задача автоматического создания онтологий на основе текстов естественного языка сегодня находится в центре внимания, как разработчиков приложений, так и исследователей в области искусственного интеллекта и Text Mining. Это свидетельствует одновременно и о ее нерешенности, и об актуальности.

При создании онтологий можно выделить несколько этапов, каждый из которых представляет собой более узкую задачу. Степень зрелости методологических и технических решений каждого этапа также различна и зависит от конечной цели разработчиков.

На основе анализа работ можно сформулировать такие тенденции, которые можно считать уже сложившимися при построении онтологий на

основе текстов естественного языка: использование в качестве исходной базы уже существующих онтологий или таксономий, типа Word Net; создание простых и устойчивых структур, например, деревьев небольшой глубины; постепенное накопление экспериментальной базы, что позволит улучшить настройку онтологии.

Данная работа имеет целью обобщение и качественное сравнение основных методов создания онтологий. В перспективе авторы надеются представить результаты собственных исследований.

**Список литературы**: **1.** *Maedche A., Staab A.* Mining Ontologies from Text. In Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management table of contents, pages pp. 189-202, London, UK, 2000. Springer-Verlag. **2.** *P. Buitelaar, P. Cimiano, and B. Magnini.* Ontology Learning from Texts: An Overview. Frontiers in Artificial Intelligence and Applications, 123, 2005. **3.** *C. Fellbaum.* WordNet: An Electronic Lexical Database. The MIT Press, 1998. **4.** *Michal Laclavik1, Martin Seleng1, Emil Gatial1, Zoltan Balogh1, Ladislav Hluchy1*: Ontology based Text Annotation, 2004. **5.** *Feldman R. and Hirsh H.*, «Mining associations in text in the presence of background knowledge», in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, USA, 1996. **6.** *Ph. Resnik.* Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language// Journal of Artificial Intelligence Research 11 (1999) 95-130. **7.** *Haifa Zargayouna, Sylvie Salotti* Mesure de similarite semantique pour l'indexation de documents semi-structures: http://www-lipn.univ-paris13.fr/seminaires /AtelierRaPC/Articles/haifa.pdf. **8.** *A. Hotho and G. Stumme.* Conceptual clustering of text clusters. In Proceedings of FGML Workshop, pages 37-45, 2002. **9.** *N. Ide and V. Jean*. Word sense disambiguation : The state of the art. Computational Linguistics, 24(1). pp. 1-40, 1998. **10.** *B. Gelfand, M. Wulfekuhler, and W.F. Punch III.* Automated concept extraction from plain text. In Proceedings of the AAAI 1998 Workshop on Text Categorization, 1998. **11.** *A. Maedche and A. Staab*. Mining Ontologies from Text. In Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management table of contents, pages pp. 189-202, London, UK, 2000. **12.** *P. Buitelaar, P. Cimiano, and B. Magnini.* Ontology Learning from Texts : An Overview. Frontiers in Artificial Intelligence and Applications, 123, 2005. **13.** *B. Nguyen, M. Vazirgiannis, I. Varlamis, and M. Halkidi.* Organising Web documents into Thematic Subsets using an Ontology (thesus). In Journees AS-Web Semantique, Paris, 2002. **14.** *G. De Chalendaret and B. Grau.* SVETLAN' – A System to ClassifyWords in Context. In Proceedings of the Ontology Learning workshop (ECAI 2000), pages pp. 19-24, Berlin, Germany, 2000. **15.** *A. Hotho and G. Stumme*. Conceptual clustering of text clusters. In Proceedings of FGML Workshop, pages 37-45, 2002. **16.** *M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis.* THESUS: Organizing Web document collections based on link semantics. The International Journal on Very Large Data Bases, 12(4). pp. 320-332, 2003. **17.** *Sullivan D.* Document Warehousing and Text Mining. Techniques for Improving Business Operations, Marketing, and Sales. – Canada: John Wiley & Sons, Inc., 2001. – 542 p., 2004. **18.** *Noy N., Musen M.* Ontology Versioning as an Element of an Ontology-Management Framework. IEEE Intelligent Systems, to appear, 2003. **19.** *Maedche A., Motik B., Stojanovic L., Studer R. and Volz R.* Ontologies for Enterprise Knowledge Management. In IEEE Intelligent Systems, Vol. 18, Num. 2, pp. 26-33, 2003.

*Поступила в редколлегию 16.03.2011.*